

## Chapter 6. Performance Assessment<sup>1</sup>

By Gretchen B. Jordan and Elizabeth L. Malone<sup>2</sup>

Developing credible and appropriate measures of performance is a major challenge facing publicly funded science organizations. Current assessment techniques are not sufficient for current requirements. Performance information is needed for effective management and for demonstrating the relevance and value of the organization's work to funders and stakeholders. In addition, publicly funded science organizations in the United States have a legislative mandate to develop performance measures and to report progress on those measures each year as part of the budgeting process. The organizational literature is clear that performance measurement influences organizational behavior, for better or worse. Consequently, managers of publicly funded science have a responsibility to participate in the development and use of better performance assessment techniques and to use the results of these improved assessments to refine their management of the public science system.

In a broad sense, the function of managers is to manage organizational performance, that is, to manage all aspects related to the setting and achievement of organizational goals and strategy. To do this, they need to establish expectations, in collaboration with their employees and stakeholders, agree upon ways to measure performance relative to those expectations, measure performance over time, and use those measures to provide feedback and take action. A variety of terms are used to describe this process and its components. We will use the term *performance assessment*.

Public agencies that conduct research are examining their assessment infrastructure to ensure that it is adequate to document and report progress and success, demonstrate their determination to improve performance, and provide accountability (Feller 2000; Rip 2000). However, numerous challenges are being encountered. This review begins by examining the drivers and difficulties in assessing R&D performance in general and in publicly funded scientific research in particular. It then discusses frameworks for measurement and provides brief descriptions of major methods and specific measures that have been found most pertinent to the assessment of performance by publicly funded science organizations. Two examples are included to illustrate both measures and methods, and the use of performance information. The review concludes with a list of areas for action and further research.

### ***The Drivers of Performance Assessment are Leading to an Increased Emphasis on Prediction***

The demand for comprehensive performance assessment of public research and technology development (R&D) programs is part of a worldwide demand for more accountability in all public programs. The most recognizable manifestation of this demand in the United States is the Government Performance and Results Act of 1993 (GPRA) (U.S. Congress 1993), which requires strategic plans, performance plans, and annual performance reports. What is driving this demand? An Organization for Economic Cooperation and Development study (OECD 1997)

---

<sup>1</sup> Related chapters include: Strategy; Change Management; and Innovation

<sup>2</sup> Chris Cluett of Battelle Memorial Institute assisted with the benchmarking discussion

concluded that three concerns are noticeable in all OECD performance management, although to different degrees: public concern that governments (1) improve performance, (2) clarify responsibilities and control, and (3) realize cost savings. Irwin Feller (2000), evaluator and advisor to R&D agencies and programs cites three other reasons for the focus on performance management in research programs: (1) a need for accountability in times of flat budgets; (2) concern that public funding is being provided in areas that do not obviously contribute to national well-being, such as civilian technologies; and (3) eroding political acceptance of claims by the research community, particularly by academics, to be self-policing.

Both common types of assessment -- measurement and evaluation -- are part of the fabric of science programs. It is the scientific method to review and improve. However, as noted in the final report of the European Union Network on Advanced Science and Technology Policy and Planning (Rip 2000):

The major rationale for evaluations has shifted and evolved from a desire to legitimate past actions and demonstrate accountability to a need to improve understanding and inform future actions.

Correspondingly, the issue of evaluation has broadened away from a narrow focus on quality, economy, efficiency and effectiveness, towards a more all-encompassing concern with performance improvement and strategy development (Rip 2000).

Georghiou and Roessner (2000) agree that organizations are increasingly called upon to provide predictive assessments prior to or during program implementation to define or refine strategies for accomplishing goals. This represents a shift from the past when program evaluations focused on retrospective or “summative” studies after projects were completed to determine whether or not objectives were accomplished and to clarify why the observed outcomes had occurred. These summative studies were used to demonstrate effectiveness and success to funding bodies. “Formative” evaluations, which look at the issues of design and implementation while the program is still underway, are more oriented to providing managers with the information they need to improve on-going programs and develop more effective plans for the future. This increasing emphasis on formative evaluations is also reflected in the U.S. National Science and Technology Council report on *Assessing Fundamental Science* (NSTC 1996), which concluded that the nature of R&D is such that “science agencies must devise assessment strategies ... designed to ... respond to surprises, pursue detours, and revise program agendas in response to new scientific information and technical opportunities essential to the future well-being of our people.”

A study conducted by the National Partnership for Reinventing Government (NPR 1999) on “best practices” found that organizations were using performance information to help make resource allocation decisions, guide staff evaluations, identify why performance and goals don’t match, determine if and how processes or goals need to be changed, and find ways of improving the measurement approach.

A key driver in this process has been the Government Performance Results Act (GPRA), which has had a major effect on performance assessment in U.S. publicly funded science organizations. GPRA reflects U.S. Congressional assumptions that agencies that establish management commitment, balance the needs of all involved and affected, and invest sufficient resources in the performance assessment process will accrue benefits (increased public confidence, better decision-making, and more informed policymaking and resource allocation) that will outweigh the costs of collecting, analyzing, reporting, and using performance data (U.S. Congress 1993).

Cozzens (1999) concludes that the new accountability that GPRA will provide could allow agencies to publicly highlight their achievements, communicate more effectively with regulators and the Federal Government, and directly address the concerns and needs of their stakeholders.

GPRA establishes a series of planning and reporting requirements that reinforced the use of prediction-oriented performance assessment:

- ◆ GPRA requires U.S. Federal agencies to use *strategic plans* to align their organization and budget structure with their mission and objectives. Strategic plans set out the long-term programmatic, policy, and management goals of the agency by outlining planned accomplishments and the schedule for their implementation. To create strategic plans, agencies must comprehensively determine, evaluate, and outline their overall mission and objectives as well as the goals and priorities of their individual programs.
- ◆ A strategic plan's goals and objectives set the framework for developing *annual performance plans*. These annual performance plans are to set out measurable goals that define what will be accomplished during a fiscal year. The goals should represent a level of accomplishment that corresponds to the resources requested and funded. Annual plans are to include:
  - The performance goals and indicators for the fiscal year
  - A description of the processes, skills, technology, and the human, capital information, or other resources that will be needed to meet the performance goals
  - A description of the means that will be used to verify and validate measured values.
- ◆ The *annual report* outlines agency performance in relation to the goals and objectives stated in their strategic and annual plans. Progress is to be connected to annual budget requests in order for policymakers and the public to have up-to-date information on how agencies use their allocated resources (GAO 1996, OMB 1999). The annual report must also discuss the results of any program assessment conducted during the fiscal year and must state the quality of the performance data being reported, including a description of the methods used to verify and validate the data (GAO 1999). GPRA requires (OMB 1999) that annual reports contain:
  - A comparison of actual and projected performance for the fiscal year
  - Justification for goals that were not met
  - A description of methods for meeting these goals in the future
  - An evaluation of the annual performance plan
  - Trend data that tracks agency performance for at least four fiscal years (when available).
- ◆ GPRA requires that all performance goals and indicators be objective and quantifiable. Agencies that are not able to define objective, quantifiable performance goals for their programs have the option of using alternative formats, if approved by the Office of Management and Budget (OMB).<sup>3</sup> OMB (1999) requires that these alternative formats be either

---

<sup>3</sup> The National Science Foundation (NSF) was authorized by OMB to use an alternative format for its performance goals. Because NSF research and education program areas mainly conduct fundamental research, and fundamental research results are difficult to quantify, NSF uses external expert review panels to qualitatively assess the research outcome goals that are found in its performance plans and report (NSF 2000). NSF combines this qualitative approach with quantitative performance goals that measure performance related to its internal investment and management process (GAO 2000).

- Separate, descriptive statements of a minimally effective program and a successful program, expressed with sufficient precision to allow for an accurate, independent determination of whether the actual performance meets the criteria of the description, or
- Some alternative that allows an accurate, independent determination to be made of how actual performance compares to the goal as stated.

### ***The Difficulties in Assessing R&D Performance as Required by GPRA***

Evaluators and study groups have found that assessing R&D performance is difficult because of specific qualities that are inherent to research and the scientific process. The difficulties of assessing R&D performance are exacerbated when R&D are funded by federal agencies as public goods. Cozzens (1999), Rip (2000), and others describe the following factors that pose particular challenges for assessing the performance of government-funded R&D and meeting the requirements of GPRA:

- ◆ *Significant research events occur unpredictably* and cannot be subject to schedules. No one can predict what discoveries will be made as a result of R&D activities because they are neither routine nor do they have specific outputs. Any goal that was predicted would likely shift before the time to evaluate the goal had arrived. And the time span is uncertain—often decades or longer—before the medical, technical, or other social benefits from R&D are fully realized.
- ◆ *Unique contributions are hard to document* because many sources of funding and contribution are often integrated in a single research program. Innovative contributions to social and economic well-being are typically the result of the national R&D system rather than any one component. The institutional landscape of national R&D systems includes contributions from academic institutions, large and small public laboratories, R&D stimulation programs, and research centers. It also includes the private sector and the international scene.
- ◆ *Less important attributes of research are the ones that can be tracked and measured.* Quantifiable outputs such as patents, publications, staff growth, and follow-on support can be only a very limited characterization of a program focused on basic research. Alles (1991) and others observe that researchers will produce large quantities of anything used to measure R&D effectiveness once they are aware of the measurement system, regardless of the relevance to good science or technology development. There is a risk of researchers falling back on safe, short-term projects with easily identified and reported outputs, for example.
- ◆ *There is no easy, accurate method to objectively evaluate fundamental research quality or result.* Peer review is the method most often used and recommended, but it is too resource intensive and time consuming to be used for ongoing or annual monitoring and has other challenges, discussed below.
- ◆ *Institutional challenges* include vested interests in maintaining or changing the system; increasing involvement and interference of a wider group of stakeholders; and questions of expertise, breadth, and representation on peer review panels (Rip 2000).
- ◆ *Communicating science* and scientific accomplishments in terms that laymen understand has always been a challenge. The increasing emphasis on making the impacts of science visible accentuates this challenge.

- ◆ *Data quality problems* exist with all methods. For example, reuse of archival data and the reconstruction of data from several years past will always be subject to data quality concerns; various quantitative measurement systems for return on investment that have been constructed are both rough and unreliable.

## ***Establishing Effective Performance Assessment Systems***

### *Creating the Organizational Infrastructure to Support Performance Assessment*

Whether an organization is redesigning its existing system for developing and using performance measurement and evaluation information or creating a new system, the same basic infrastructure is required. Without an infrastructure, lack of focus, low levels of participation, disagreement, or incomplete implementation will limit the performance management effort. The NPR (1997, 1999), GAO (1996, 1999), Canadian Federal Government (1994), M.G. Brown (1996) and others suggest that an organization must have the following for a performance management system to succeed:

- ◆ Leadership commitment
- ◆ A desire for accountability
- ◆ A conceptual framework
- ◆ Strategic alignment
- ◆ Knowledgeable and trained staff members
- ◆ Effective internal and external communication
- ◆ A positive not punitive culture
- ◆ Rewards linked to performance
- ◆ Effective data processing systems
- ◆ A commitment to and plan for using performance information.

Use of the performance information is identified as one of the critical indicators of a successful organizational performance management program (Hatry 1999). To be successful, a performance measurement system must generate data that are used to enhance management practices and achieve goals, provide accountability for all of the stated goals, improve performance, allocate resources, and make informed policy decisions. The National Partnership for Reinventing Government (NPR 1999) stressed the importance of using the results of performance measurement to involve employees and customers in the performance processes (see also Chapter 10, “Participative Management and Employee and Stakeholder Involvement”). *The American National Standard Quality Guidelines for Research* published in 2000 by The American Society for Quality (a leading quality improvement organization with more than 100,000 individual and 1,100 corporate members worldwide) also emphasized the importance of using performance assessments to understand and improve programs and to identify, correct, and prevent problems.

### *Frameworks for Assessment*

What to measure depends upon the information needs of stakeholders and the nature and goals of the organization or program and its context. Best practice suggests that choosing what to measure begins with defining an explicit model or conceptual framework that describes what the program or organization’s purpose and goals are and traces the pathways from R&D process to outputs

and potential outcomes and impacts. Such a framework, often called a “logic model,” clarifies assumptions and helps internal and external stakeholders develop a shared understanding of performance expectations. The model should attempt to capture the complex and unpredictable character of R&D and the multiple objectives of different aspects of performance. Using a framework can provide a means for matching an organization’s basic characteristics and circumstances with assessment strategies, and help choose effectiveness criteria that recognize views of multiple stakeholders. The framework, whether explicit or implicit, serves as the basis for the following decisions about assessment (Georghiou and Roessner 2000).

- ◆ The choice of what to measure
- ◆ How and when to measure
- ◆ How to interpret the results.

In other words, the framework provides the basis for developing the research design for the assessment. A research design is the logic that links data collection and conclusions to the initial questions of the assessment. The research design is a blueprint of what questions to study, what data are relevant, what data to collect, and how to analyze the results (Yin 1984:18-20). It should specify (Yin 1984):

- ◆ The purpose and scope of the assessment
- ◆ The principal questions to be answered
- ◆ The audience
- ◆ Its propositions, if any
- ◆ The unit of study
- ◆ The time frame
- ◆ The logic linking the data to the propositions
- ◆ The principal methods of data collection and analysis to be used
- ◆ The criteria for interpreting the findings.

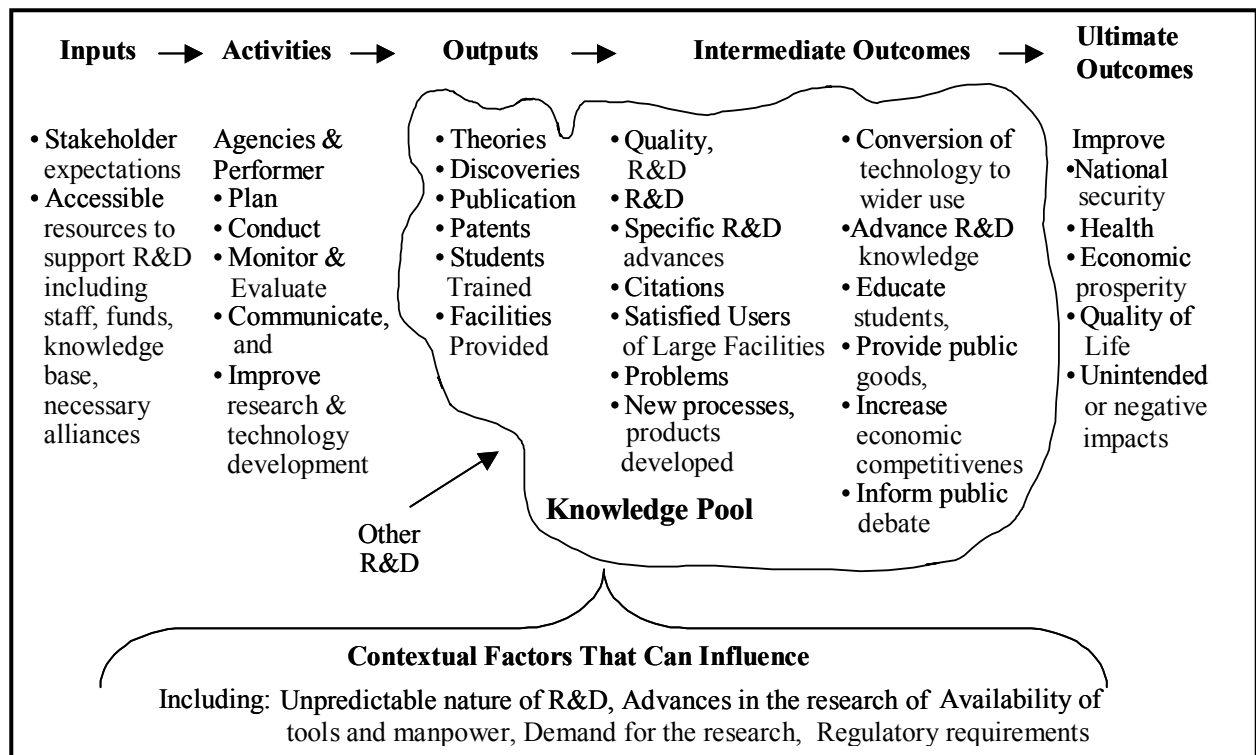
An important part of developing a framework for assessment is to put a program in the context of its contribution to broader goals. The elements and flow of the framework depend upon the type and purpose of the research. A framework for a fundamental research organization will differ from that for an applied research or technology development program. Expected outcomes will differ depending upon the primary purpose or mission of the organization. Laredo and Mustar (2000) characterize five “strategic profiles” for a research organization, most often seen in combination with one another: (1) advance knowledge, (2) train students, (3) provide public goods, (4) provide economic advantage, and (5) inform public debate.

The framework should also show or link to agency strategic plans and, for public programs, stated Federal objectives. The overall strategic goals for the long-term stewardship of public science were outlined for the U.S. Federal Government in the report “Science in the National Interest” (1994). These goals include:

- ◆ Maintain leadership across the frontiers of scientific knowledge
- ◆ Enhance connections between fundamental research and national goals (improved health, environment, prosperity, national security and quality of life)
- ◆ Stimulate partnerships that promote investments in fundamental science and engineering and effectively use physical, human, and financial resources
- ◆ Produce the finest scientists and engineers for the twenty-first century
- ◆ Raise the scientific and technological literacy of all Americans.

A simple program logic model for the R&D process is shown in Figure 2. A logic model describes Inputs, Activities, Outputs, a logical sequence of Outcomes, and external contextual for a program or organization and the logical linkages among these. What is the program doing, for whom, and why? This logic model uses the notion of “knowledge pool” from Cozzens (1997) to highlight the difficulty of tracing the transfer of specific R&D discoveries to more applied research, development, and conversion to wider use.

Many organizations and researchers use a logical framework to design assessments without using this terminology. House and Schull (1985) advocate a system-wide evaluation process that encompasses all segments of a laboratory’s operation. A Pacific Northwest National Laboratory report (1992) that looked at measuring Laboratory-directed R&D projects used both quantitative measures (intellectual property, publications, human resources, and follow-on support) and qualitative outcomes (tracing individual “success stories”). Industrial firms with research and development (R&D) components have used a framework approach when choosing what to measure. For example, Werner and Souder (1997), after reviewing the methods for measuring R&D performance available from 1956 to 1995, suggest an approach that includes an index of past performance, an index of on-time project completion, a future potential index, and a peer rating audit. They conclude that integrated approaches that combine qualitative and quantitative metrics are more accurate, reliable, effective, and flexible than individual metrics, although they require a larger investment of time and resources to develop and implement. Werner and Souder present guidelines based on an organization’s need for comprehensiveness, data that are available, type of R&D to be measured, and resources available for measurement.



**Figure 2. The Logic of R&D Program Performance**

## *Constructing Performance Measures*

Considerable advice is available on developing and choosing measures or indicators of performance. GAO (1996, 1999b), the OMB budget submission guidelines (OMB 1999) and the administration's National Performance Review (NPR 1997, 1999) provide guidelines for choosing measures for GPRA. Mark Graham Brown's book (1996) on measurement also outlines the characteristics that make up good measures and the measurement attributes that produce high-quality performance data. The GAO (1997) report on research indicators and the COSEPUP (1999) study on evaluating research discuss what aspects of research should be measured. The Auditor General of Canada (Canada 1994) is another good source of guidance.

Summarizing across these sources, the set of key performance indicators or measures should:

- ♦ Be centered on evaluating a program or activity's core purpose; drive the organization to take appropriate actions
- ♦ Be robust enough to respond to the multiple priorities that make up a program or organization
- ♦ Be accepted and meaningful to the organization and its stakeholders
- ♦ Be outcome or output oriented; evaluate actual progress toward stated goals and demonstrate results
- ♦ Be drawn from and related to projected resources and the budget process
- ♦ Have an "appropriate" timeframe
- ♦ Produce objective, clear, and (usually) quantifiable results.

The typical areas of measurement are listed in the logic model in Figure 2. For an extensive discussion of measures see Geisler's book on science and technology metrics (Geisler 2000). The brief description of some commonly used measures that follows is organized into the three areas required by GPRA: inputs, capabilities, and process measures, or in GPRA language, means and strategies; outputs; and outcomes.

### **Typical Inputs, Capabilities, and Process Measures**

This group of measures examines aspects of resources and processes for transforming resources into outputs, including management processes. One primary indicator of how much research is being performed is, of course, the amount of money spent on R&D. These data have been refined over many years and are generally available. This data is the basis for the "cost" side for the organization in benefit-cost analysis. It also is the basis for showing research intensity, a comparison of expenditures to total firm expenditures or to a country's gross domestic product. The premise that higher expenditures indicate higher returns or results is not generally accepted, however. "The level of spending is not a reliable indicator of the level of results achieved by research" (GAO 1997). Other commonly used input measures are measures of the number and quality of staff available to an R&D organization. Capability measures also include measures of capital investment in facilities and equipment, the knowledge base, and core competencies.

Input measures may also relate to the management and conduct of the scientific research. Demonstrating good management of science helps stakeholders trust that good outcomes are likely to be forthcoming in addition to demonstrating fiscal responsibility. The OMB requires all science agencies to have three measures of process: the percentage of projects that are peer reviewed, the downtime for scientific facilities compared to scheduled downtime, and cost and



schedule for facility construction and upgrade. Process measures are particularly important for management to improve. Other measures of process are employee retention and employee satisfaction and indices of the organizational culture. Aspects of the conduct of R&D that could also be measured are the quality of the supporting infrastructure, communication among teams, and degree to which scientists are unhindered in their research but subject to clear ethical and other restrictions. A book by Ellis (1997) looks exclusively at R&D process measures based on his experience with firms in the Industrial Research Institute.

## Typical Output Measures

An output measure is the tabulation, calculation, or recording of activity or effort and can be expressed in a quantitative or qualitative manner. Output measures have the advantage over outcome measures in that they are directly observable in the shorter run. They provide evidence that developments are underway that are expected to lead to ultimate goals. Often the link is plausible on its face, such as improvements in medical science or technology being linked to improved health of citizens. In other cases, experience or evaluation studies have empirically shown a link between past outputs and outcomes. Experiences from pilot efforts under the Government Performance and Results Act have reinforced the finding that output measures are highly specific to the management and mission of each Federal agency and that no single indicator exists to measure the results of research (GAO 1997). Because companies are profit oriented, many of the indicators tracked by the private sector cannot be directly applied to the Federal government.

A primary measure of output is the quantity of knowledge produced. This is often measured by publication counts such as the number of peer-reviewed publications. This reflects the generation of knowledge and the amount of information presented to the scientific community and the public. Invitations to speak and present at conferences and positions in professional societies are also output measures, as are awards and prizes. Other output measures are the number of patents, devices, and software developed. Technical milestones and progress toward reaching them can be a credible output measure, particularly if these milestones were judged “ambitious” during an expert review. In the area of human resource development a typical measure is the number of students trained. And where an organization provides scientific user facilities, the number of users and their satisfaction with scientific facilities are reported. Customer or user evaluations can also be structured to gather data on use made of the outputs, thus they can provide information on short-term outcomes.

## Typical Outcome Measures

An outcome measure is an assessment of the results of a program activity compared to its intended purpose and the science itself. As shown in the logic model in Figure 2, there is a sequence of outcomes, following a dynamic, complex, and unpredictable R&D process that can be summarized as the creation of a new idea, development of that idea, and conversion into wider use. GPRA and the more general demand for accountability prefer that the outcomes measured are “ultimate” outcomes, that is, tangible impacts on society or the economy linked to broad national goals, such as new products or processes. The impacts may be “potential” rather than actual and important because they are “options” on future tangible benefits. Benefits can also be advancing of knowledge, informing the public debate, and educating scientists and engineers.

For fundamental science, tangible outcomes can be assessed retrospectively, linking funding of research to a research discovery that led to a cleaner environment, for example. Economic studies and case studies are used to assess impacts. If costs are also determined, return on investment and cost benefit determinations can be made.

For current research, the outcome measures recommended by COSEPUP and other studies are expert judgment on the quality of the R&D, the relevance to science and important national goals, and scientific leadership, that is comparison of quality and capabilities to other agencies or countries. The expert judgment often reviews quantitative measures such as citations of publications in journals and patent applications, in addition to the opinions of the peers. Network analysis and other new methods are being developed to demonstrate advances in knowledge. R&D performers, and their managers are trying to do a better job of describing the potential “next stage” uses of their research and come up with investment criteria with which to choose between projects. More information on these measures is found in the discussion of methods that follows. A review of the economic methods for assessing fundamental science can be found in Popper (1995). The Bozeman and Melkers book (1993) on assessing R&D impacts covers a variety of methods.

### *Selecting Appropriate Data Collection and Analysis Methodologies*

Collecting performance data is the next step once an organization has defined its performance measures and goals and has its performance management infrastructure and plan. To be effective, data collection must be followed with accurate analysis, feedback to all organizational levels, and use of the performance information to improve performance. Because of the high cost and level of effort associated with measuring performance and the great potential benefits, organizations must strive to collect high quality data and to make use of the information for positive organizational change (GAO 1996; NPR 1997, 1999; M.G. Brown 1996; Hatry 1999). The GAO (1999b) discusses some of the methods agencies have implemented to ensure that their performance data are valid and meaningful. The GAO groups the methods into four general categories that include actions for senior management, individual program managers, and technical staff:

- ◆ Establish a commitment and capacity for data quality.
- ◆ Assess the quality of existing data.
- ◆ Respond to data limitations.
- ◆ Build quality into the development of performance data.

Measures and methods are often classified as quantitative or qualitative, and as objective or subjective. Even inherently qualitative measures may be quantified. Werner and Souder (1997) point this out and talk of four types of metrics: quantitative objective, quantitative subjective, qualitative, and combined. Quantitative objective metrics are based on numerical measures of R&D input and output, for example, staff count, R&D cost, time spent, number of publications and/or citations, number of patents, cost reductions, and goals met. Quantitative subjective metrics are based on non-numerical judgments that are converted into numeric values and ratings via profiles, scaling models, checklists, and scoring models. Qualitative metrics measure human resource and other aspects of R&D performance using self-assessment, supervisory rating, peer rating, and external audit.

The methods chosen to do the measurement and evaluation of performance will depend on several things. Some methods are more appropriate at some times than others. Building on a

study of methods for assessing the socioeconomic impact of R&D conducted by the Federal Government of Canada (Canada 1993), choosing methods depends on six factors:

- ♦ Whether the assessment is occurring before or after the R&D is completed
- ♦ The type of R&D involved (basic/strategic research, applied research, product/process development)
- ♦ The purpose/category of the R&D (advance knowledge and educate students, inform policy, provide public goods, or support industry)
- ♦ Whether the assessment is of outcomes or impacts of R&D or the quality, activities, or other aspects of the research process
- ♦ The unit of analysis used, that is, whether assessing the aggregate impact of all science on the economy or impact of individual programs or projects
- ♦ Resources available for measuring and evaluation.

### Established Methods Frequently Used to Assess Performance in Science-Oriented Organizations

The choice of methods and analytic tools for performance assessment is determined by the nature of the events being studied, the questions to be answered, and the nature of the available data. The particular capabilities, experience, resources, and preferences of those designing the study also influence method selection. (For a more detailed discussion of the various methods used in assessing the performance of science organizations, see Branch et al. 2001.)

*Peer and Expert Reviews.* Peer and expert reviews make use of the opinions and judgments of recognized experts in particular fields to evaluate publications, individual research projects, research programs, organizations, and fields of research. Kostoff (1997) defines a peer as a person with expertise in the specific or allied technical area of the research being reviewed or in technology, systems, and operational areas that may be impacted in the future by the research being reviewed. Peer reviewers can be outside researchers with knowledge of the research area, research managers, or professional evaluators within or outside the research organization sponsoring the science. Peer review is the primary method used by Federal agencies to assess the value of their research activities. Peer review can be used at many stages of the research process: proposal selection, project and program evaluation, and evaluation of R&D impacts.

Kostoff states that “a properly conducted research program peer review can provide credible indication to the research sponsors of program quality, program relevance, management quality, and appropriateness of direction.” Several studies concluded that peer review is the most effective method for evaluating the quality and value of fundamental research, whether the research is completed, on-going, or in the future (Canada 1993, COSEPUP 1999, Bozeman 1993). Using the peer review method, experts have the ability to discuss the research with the investigators, assess the methodology, and compare the research to what they know from experience has either succeeded or failed in the past. This process should result in a consensus on the quality of the research.

Peers evaluating the quality of projects is the most common type of expert review, but there are others. A type of peer review, advisory committees operate at the level of a program or a science-sponsoring organization. They evaluate performance against goals, comment on directions for the future, and give advice on management issues. Two types of review typically conducted by experts who are not necessarily peers in the specific science field are relevance review and benchmarking (COSEPUP 1999). A relevance review is intended to assess whether agency research programs are consonant with its mission.

For example, if a goal of DOE is to produce cheaper solar energy, it is consistent with the agency's mission to understanding the physical properties that determine the ability of materials to convert solar radiation into electrical energy. A careful relevance review could indicate the most promising directions for future research, both basic and applied (COSEPUP 1999).

The quality of peer review is dependent on how the review is organized, the expertise and competence of the reviewers, and the resources available for the review. The most significant limitation of peer review is that it is based on the subjective judgments of individuals and thus prone to bias. The bias may stem from many sources, including favoritism, the researcher's past history and experience, the size of the institution supporting the research, and the fact that the researcher may be not only a colleague but also a competitor (Bozeman 1993, Canada 1993, Callon et al. 1997, Chubin 1994, Kostoff 1997, COSEPUP 1999). Peer review is also of limited use in fields with low rates of publication. Unless the manager of the review is well versed in the field's research activity, peers may be hard to identify. Also in fields with low publication rates, researchers often produce proprietary information and are less willing to present their data for review by the competition (Bozeman 1993, Chubin 1994). Another limitation of peer review is that it can be expensive to successfully organize and implement. Time, resources, and effort are required to screen and recruit peers, conduct site visits, and interview researchers (Bozeman 1993, GAO 1997, Kostoff 1997).

*Key Indicators.* Indicators point towards, suggest, or provide evidence of a likely outcome. Indicators are used when it is too costly or difficult to get at actual measures of outcomes. Indicators are often used as interim evaluation results for research projects that are currently underway. Often indicators are presented as annual or cumulative trend data, such as the number of annual awards received from outside organizations, or total number of graduate students supported. Awards are an indicator of significant contributions to the scientific knowledge base. The number of graduate students supported is an indicator of the future workforce and contribution to capabilities in the field.

The data used for indicators often comes from application of the other methods of evaluation, such as survey, bibliometrics, and statistical methods. Program documents and observation are other common sources. Indicators are usually in a set balanced across the perspectives and needs of stakeholder groups, such as in the "Balanced Scorecard" approach of Kaplan and Norton (1996). The Balanced Scorecard has indicators from the perspectives of Finance (profit or success for industry), Customer, Employee, and Operations. Another approach to a set of indicators is "converging partial indicators," a group of indicators that measure a particular aspect of performance. Indicators are viewed as objective, trendable, and quickly and fairly easily understood. Not all are objective, however, and they can be easy to manipulate in terms of providing, analyzing, and displaying the data. Most studies suggest that indicators cannot be used alone without explanatory information and qualitative data.

An example of a set of key indicators is the a menu of 50 metrics in the Technology Value Pyramid developed by The Industrial Research Institute (IRI) to assess and predict R&D performance (Tipping et al. 1995). The 280 IRI members, companies that carry out over 80 percent of the industrial research effort in the U.S., cooperatively developed the pyramid, which has three levels of measures. The Foundations level measures asset value of technology and R&D processes. The Strategy level measures portfolio assessment and integration with business. And the Outcomes level measures value creation.

*Economic Methods - Return on Investment.* Estimates of the social or economic benefits that organizations receive from initial investments in R&D can help demonstrate the value of the R&D as well as help organizations make better decisions about budget and human resource allocation. The results of economic methodologies typically indicate that R&D activities produce high overall rates of return (GAO 1997). Assessment methods that are available to measure many of the economic facets of an organization's research include return on investment, the production function, customer surplus, and increased benefit to industry and society (Averch 1994, Cozzens et al. 1994, NSTC 1996, Tassey 1996, GAO 1997, Geisler 1999, 2000). These methods work best for applied research and for aggregate assessments of the influence of all fundamental research on society (Cozzens et al. 1994).

Economic methods require discrete values for research input and output for use in mathematical functions, models, and equations. However, the appropriate data may be uncollected, unreliable, or proprietary (Averch 1994, Cozzens et al. 1994). Furthermore, the factors involved in derivation of economic benefit are complex. Using a single value to represent the economic, social, technological, and behavioral issues may distort the importance of each individual factor (Cozzens et al. 1994, GAO 1997, Geisler 1999). Finally, economic calculations are hindered by the long time period between initial R&D investment and final realization of benefit (Tassey 1996, GAO 1997, COSEPUP 1999, Geisler 1999). By the time benefit can be determined, isolation of the inputs that led to the benefit may be difficult or impossible. Economic measures have only a tenuous relationship to Federally sponsored R&D, since improving productivity or producing an economic return is less and less the primary *a priori* justification for these programs (GAO 1997, OTA 1986).

*Case Studies.* The case study is a research method that extensively describes and analyzes complex situations, in context, to answer questions such as efficiency and impact. Bozeman and Klein (1999) state that the goal of case studies in assessing R&D impact typically is to answer two questions: "What are the linkages between R&D and economic growth?" and "Are R&D projects meeting the policy objectives established for the organizations that mandate addressing linkages between R&D and the economy?" Case studies are also used to assess the development and outcomes of individual programs and projects. Retrospective case studies are "historical accounts of the social and intellectual developments that led to key events in science or applications of science" (COSEPUP 1999). Project Hindsight (Sherwin and Isenson 1967), sponsored by the Department of Defense, and Technology in Retrospect and Critical Events in Science (TRACES) (IIT 1968), sponsored by the National Science Foundation, are two examples of very large-scale case studies. These two projects traced discrete technological advances backward to isolate the specific activities and events that led to their development.

The most significant benefit of case studies is their comprehensiveness. Case studies often encompass the use of several other evaluation methods such as site visits, surveys and interviews, and content analysis. They look at a specific situation in detail. A limitation associated with case studies, however, is their lack of structure, which allows the incorporation of biased perspectives of the individuals that are interviewed for the foundational information (Kingsley 1993). Also, results typically vary according to who is conducting the assessment, which researcher is interviewed, and how long it has been since the research was conducted. Third, depending on the scope of the project, it is also possible for a case study to require a large investment of time and effort (Kingsley 1993).

*Bibliometrics.* Bibliometrics and citation analysis analyze databases of publications and patent filings to assess the quantity, quality, significance, dissemination, and intellectual linkages of

research. Analysis may be for an individual, program, or discipline. The NSF's *Science and Engineering Indicators* contain many bibliometric indicators at the national level. The methods are also used to measure the progress, dynamics, and evolution of scientific disciplines. This analysis is particularly useful for assessing basic research because a typical output of basic research is publication. Patent-citation analysis, in contrast, is generally more important as a tool for assessing technology. Both the bibliometrics and citations methods are often used in combination with other methods, such as with the historical tracing, peer review, and classical case study methods.

Advantages of the method are that the data captures a principal output of basic research programs and thus provides a means of assessing the quantity, significance, trends, and linkages. Data is fairly easy to collect, relatively objective, and easily understood by diverse audiences. Bibliometrics is generally accepted as a valid form of assessment. Disadvantages include concern that counts indicate quantity of output, not quality. Normalization approaches may not adequately adjust for differences in the quality and importance of journals, nor for differences in publication practices across organizations and disciplines. In particular, Fuller (1997) provides a critique of the Science Citation Index, as an equivocal and unreliable measure because articles can be cited for many reasons and because most scientists, to avoid the "morass" of journal articles, circulate prepublication manuscripts among themselves to keep abreast of leading-edge work in their fields.

*International Benchmarking.* COSEPUP (2000) identified international benchmarking as a promising method for comparing research and development programs in an objective, timely, and cost-effective manner that could produce consistent results about the quality of research despite differences in reviewers, methods, and objectives. Benchmarking is defined as:

a formalized quality process that is used to continuously measure products, services, processes, and practices against competitors or "best practice" companies, determining how the best-in-class companies achieve those performance levels and applying that knowledge to your own operations to achieve competitive advantage (Ransley 1994).

Benchmarking is a fairly recent management strategy, pioneered in the late 1970s by Xerox Corporation. Xerox was beset by outside competition, both national and international, and the company was not only stagnating, it was falling behind, at risk of financial collapse. Xerox embarked on a process of careful assessment of its competitors to learn how they were able to produce and sell copiers more efficiently and at less cost than Xerox was doing. With this knowledge, Xerox managers focused on making substantial changes and improvements in how they conducted each of the critical aspects of their business. By committing to learning from benchmarking and applying what they learned to changing their business practices, they dramatically turned around the fortunes of their company.

Benchmarking has been successfully used in industry but government experience with this method is very limited. Benchmarking methods can be used on several different scales to compare the research activities of a process, a metric, a program, organization, industry, or country (Ransley 1994, COSEPUP 1999, 2000). Benchmarking takes time, attention, and resources, so it is usually reserved for a change that senior management has determined is necessary to meet strategic objectives. This is true even for benchmarking related to metrics. For example, Raytheon collects benchmarks for a few corporate metrics before beginning to measure so that it knows what level of performance is reasonable to achieve (Raytheon 1998, n.d.).

To benefit from benchmarking, the organization must want to commit itself to change and recognize that success at benchmarking requires a concerted effort over a period of time (Anderson and Pettersen 1996; Camp 1995; 1998). A benchmarking team established to implement the process is also foundational to the steps in the benchmarking process summarized here:

- ♦ The first step is to decide what is important to benchmark. What are critical success factors for achieving the organization's mission, and which processes and attributes are most critical to that success? Benchmarking centers on aspects of these critical success factors.
- ♦ Next the organization must find appropriate benchmarking partners. The search can be internal or external, among either similar or very different organizations that may excel at a particular process. Whichever partners are selected need to agree to participate in the study.
- ♦ Then data is collected via documents, interviews and surveys, and site visits. Members of the benchmarking team observe and document the partners' processes and attributes, both performance and practice. What are they doing that makes them better than anyone else at this particular process? At this stage it is particularly important to be sensitive to potential legal and ethical issues (Bendall et al. 1997).
- ♦ Analysis is done on the gap between performance in the chosen area and the benchmarking partner's observed performance. What has caused the organization to fall behind the partner's performance in a particular area, and how can it not only catch up but surpass the partner?
- ♦ The organization must then communicate the findings from the analysis and gain organizational acceptance for the required changes. An implementation plan will need to be designed and put into action. Performance measures need to be developed that will provide management evidence that change has or has not occurred.

The benchmarking method recommended by the COSEPUP study relies on a select panel who provide expert judgment based on their knowledge supplemented with quantitative and qualitative data collected using other methods. Experts selected for the panel need to have broad knowledge of the field. The most effective panels have a balance of expertise in the field, related fields, and users of the research, and have a reasonable level of participation from experts from other countries. COSEPUP suggests asking the expert panel members to answer the following types of questions:

- ♦ What is the position of the organization's science relative to science elsewhere in U.S. and in other regions and countries in the world; i.e., who is currently achieving the most promising advances?
- ♦ What will be the organization's scientific position in the near and longer terms, given current trends?
- ♦ What are the key factors influencing relative U.S./DOE performance in the field?

They recommend that expert judgment be informed by and supplemented with the following types of evidence (which would vary from field to field):

- ♦ Experts in the field planning a "Virtual Congress."
- ♦ Computer-assisted analysis of research project using abstracts and/or citations.
- ♦ Citation analysis, journal-publication analysis, prize analysis, and analysis of international-congress speakers.
- ♦ Quantitative data analysis (e.g., funding, where the best students are going to work).

- ♦ Other quantitative and qualitative data from NSF and similar international organizations, professional societies, and perhaps from an email questionnaire.

*Customer and User Evaluation.* Customer and user evaluation is a way for organizations to obtain feedback on their performance from both their external customers and their internal customers or employees. The customer of R&D programs cannot always be identified because outputs of new knowledge go into a general pool of knowledge. In other situations, such as the users of a large scientific facility or research that is done at the request of an individual or organization, users of the facility or research are identifiable as customers. NPR (1999) states that customer “feedback can tell an organization how well it is performing and communicating, and can also identify emerging issues and problem areas.” The Army Research Laboratory uses customer evaluation as part of its evaluation construct. The NPR team cites the practices of several high-performance organizations to obtain feedback information, including web sites, forums and meetings, focus groups, interviews, and written, mailed, or telephone surveys. Using these customer evaluation methods, organizations can obtain information on customer condition, attitude, action, or behavior after using a product/service (impact); overall satisfaction with a product/service; ratings of specific product/service quality characteristics; extent of awareness and use of a service; reasons for dissatisfaction; and suggestions for improvement and new needs. Although customer evaluations are quantitative, they are also subjective. This subjectivity limits the data by making it prone to a number of sources of bias.

## Emerging Innovative Methods

The methods for assessing R&D discussed thus far have been the traditional approaches to performance assessment. However, innovative methods for evaluating R&D are emerging, five of which are briefly described here: (1) evaluating knowledge-value alliances, (2) evaluating human resource capacity, (3) assessing organizational effectiveness in recognizing competing values, (4) use of stock option pricing theories, and (5) a multi-module approach for evaluating portfolios of programs. The U.S. Department of Energy’s (DOE) Office of Science has funded research that generated the first three of these innovative methods.

*Knowledge-Value Alliances.* The Knowledge Value Framework (Rogers and Bozeman 2000) provides a way to assess and evaluate the research process without concentrating on individual disciplines or projects. The core idea of this framework is that the significant results of research are not produced by single projects or entities but by interrelated individuals, groups, and institutions that share resources, skills, and ideas in pursuit of a common goal. The Knowledge Value Framework is made up of two concepts: the knowledge value collective and the knowledge value alliance. The knowledge value collective is “a set of individuals connected by their uses of a particular body of information for a particular type of application – the creation of knowledge.” The knowledge value alliance “is an institutional framework binding together a set of directly interacting individuals, from multiple institutions, each contributing resources in pursuit of a transcendent knowledge goal.” By orienting assessments of R&D activities based on knowledge value alliances instead of single projects, a more accurate assessment can be made of overall impact, output, and outcome.

*Measuring Human Capital.* Bozeman et al. (1999) offer a way to look at research performance based on the value of human resources, in particular, “capacity,” the ability of groups of scientists, engineers and the users of their work to grow and sustain and to make the most of the available talent reservoir. Assessing capacity requires a long-term view and longitudinal data,



both rare in social research. The authors have developed a “scientific and technical human capital construct” and illustrated the approach with a simple model for explaining scientists’ progression rates to full professor – using curriculum vita data and longitudinal analysis with event history models. Their next step will be to model the effects of grants and other public funding support on scientific and technical human capital.

*Operational Effectiveness Recognizing Competing Values.* A study sponsored by the DOE (Jordan and Streit 2000) has generated a framework and employee attitude survey for assessing the operational effectiveness of R&D organizations. The framework uses the Competing Values theory of Cameron and Quinn (1999) that captures four of the most common perspectives and models of organizational effectiveness. Each model stresses the importance of different values for creating an effective organization. For some, creativity and a structure and focus that foster innovation and cross-fertilization are most important. Some emphasize productivity and setting and achieving rational goals. Others have a structure and focus that concentrates on internal systems, bureaucracy, and efficiency. Still others see human resource development, morale, and commitment as the means to achieving quality outputs. Attributes of R&D organizational effectiveness covering all four perspectives have been defined through literature review and 15 focus groups. An employee attitude survey has been implemented in several R&D organizations, both public and private.

*Option Pricing.* Vonortas and Lacky (2000) introduce the idea of using stock option pricing theories to make decisions about investing in long-term research programs. The authors suggest that the decision to invest initially in an R&D project with a highly uncertain outcome is conditional on revisiting the decision sometime in the future and that this is similar to buying financial call and put options. For example, an initial R&D investment will permit (but not oblige) the investor to commit to a particular technical area. The authors point out that methods in stock option pricing theory have matured enough that they can be transferred to the problem of selection among highly uncertain, long-term R&D investments in the public sector. This mechanism would help justify long-term R&D investments and allow the use of more appropriate discount rates because it better differentiates among technologies in various stages of development than current methods.

*Multi-Module, Multi-Program Evaluation.* In response to the need to evaluate portfolios of programs and ultimately policies in the European Commission’s Framework programs, an assessment was designed for the Technology Development Center of Finland using a multi-module approach that involves combinations of traditional assessment methodologies (Guy 2000). This study of eleven R&D programs assessed program relevance, efficiency, quality, effectiveness, and strategy. To gain insight into the Technology Development Center’s portfolio of programs, the study team also evaluated the appropriateness of the program mix. The results of separate assessment modules are synthesized, combining the traditional “depth” associated with the detailed assessment of published outputs (peer review) and individual projects (impact case studies) with the “breadth” of shallower but more wide-ranging reviews of complete programs (expert panel overviews).

## Case Examples – Metrics and Methods













### *The Army Research Laboratory Evaluation Construct*

The Performance Evaluation Construct developed by the Army Research Laboratory (ARL) during a pilot project for performance measurement under GPRA (E. Brown 1996) assesses the laboratory's overall functional health in terms of quality of research, relevance of research, and productivity. ARL's framework uses peer review, customer evaluation, and quantitative metrics. As shown in Figure 3, the Performance Evaluation Construct chooses evaluation methods based on what aspect of research it best evaluates. For example, peer review is the most useful technique for assessing research quality. The construct has received criticism for its lack of outcome measures.

In this approach, six panels of experts each evaluate one of the ARL business areas and conduct an annual review of the scientific and technical quality of ARL programs, the state of ARL facilities and equipment, and the preparedness of ARL technical staff. Customer evaluation is used to evaluate research relevance and productivity. A stakeholder advisory board and internal customer surveys provide information on management views, external opinion, and product quality.

A menu of 54 quantitative metrics is used to obtain a more limited assessment of relevance, productivity, and quality. Metrics available to illustrate ARL's level of performance fall into the areas of

- ♦ Preeminent in key areas of science
- ♦ Staff widely recognized as outstanding
- ♦ Seen by army users as essential to their mission
- ♦ Intellectual crossroads for the technology community
- ♦ Miscellaneous accounting and demographics.

Why...			
How...	Relevance	Productivity	Quality
• Peer Review			
• Metrics			
• Customer Evaluation			
 Very Useful	 Somewhat Useful	 Less Useful	

**Figure 3. The ARL Performance Evaluation Construct**

## *Department of Energy Science and Technology Program Assessment*

Most DOE technical programs are judged on quality and relevance of the science, effectiveness and efficiency of the management of science, and management of large user facilities:

- ♦ Quality of the science includes judgments on innovation, sustained progress and impact on the field, and recognition by the scientific community— including awards and invited talks.
- ♦ Relevance to DOE missions and national needs includes contribution to U.S. leadership in science, alignment with DOE program guidance, and interaction with other R&D programs.
- ♦ Effective and efficient research program management includes well-developed research plans, optimal use of personnel, meeting milestones, and overcoming technical problems.
- ♦ Effective construction and operation of large user facilities includes building on budget and schedule, operating to serve a diversity of scientific users, and facility dependability.

In addition to measuring technical success, each DOE laboratory assesses performance objectives, criteria, and measures in three other areas (University of California 1995):

- ♦ Operational Effectiveness includes economy of operations and productivity, focus on best business practices, and performance improvements measured over time.
- ♦ Stewardship includes managing regulatory compliance and commitments, demonstrating how responsibilities related to contractual, legal and regulatory requirements are being met, and ensuring the safety and health of the environment.
- ♦ Customer Satisfaction includes focus on improving the quality of the overall product in a manner that is timely and aligned with customer requirements.

The DOE assesses the scientific performance of its programs and laboratories primarily through peer review and self-assessment, including monitoring progress on technical milestones. DOE uses “merit review with peer evaluation,” a formal, competent and objective assessment process using specified criteria, and the review and advice of qualified peers, to guide research direction and to assess research progress. There are also assessment studies such as customer evaluations, case studies, and cost benefit analysis that document accomplishments and gather more detailed information on what worked and why.

The annual laboratory self-assessment and DOE headquarters laboratory appraisal include the results of review committees’ opinions of the quality and management of the science. The laboratories review all technical functional areas, which are groups of programs, at least every three years. Self-assessments are under the direction of the responsible laboratory line management, and include implementation of planned improvement actions and the appropriate removal of barriers to improvement. There are periodic summaries of performance and identification of key issues by functional area. Currently the links between laboratory appraisal and DOE or SC strategic plans and GPRA response are not clear.

### ***The Challenges Ahead***

The considerable challenges that lie ahead for performance measurement and management of public science have been mentioned throughout this review. The current requirements for reporting and using performance information in decision making, whether for external funders

and stakeholders or for science managers, cannot be met with the existing methods and infrastructure. Changes are needed, but these changes must be made carefully with full recognition of the complex and unpredictable nature of science and R&D and its place in the larger innovation system. It is understood that measurement perturbs any system. The objective would be to perturb it as little as possible and in the right directions. The major challenges are to:

- ♦ Develop a shared understanding with funders about the role of public R&D within the national innovation system, along with an understanding of the R&D process and timing and shared responsibility for benefits.
- ♦ Develop evaluative criteria for choosing among public R&D programs that reflect this common understanding accurately enough not to perturb the R&D activities in ways that diminish creativity and outcomes. Investment decisions are primary. Criteria used in these decisions will drive choice of measures and assessments.
- ♦ Develop “performance stories” for R&D programs at several levels of detail that communicate to the public what it is the program attempts to discover and why that is important and relevant to national goals. This will require more active involvement of scientists and engineers and work with experts in communicating technical subjects to the public.
- ♦ Develop better ways to measure and assess the various aspects of R&D performance, particularly outcomes. Existing methods can be modified and new methods can be developed. Protocols for collecting data are needed so data are reliable and valid, and where appropriate, available for decision making at higher levels of aggregation. Tools for mining and displaying data are needed.
- ♦ Establish the commitment and resources to collect, process, store, and make accessible key performance data. Data could include funding by principal investigator and end-of-project reports that include outputs and, where appropriate, outcomes. It will take years to develop the knowledge base needed to track projects through time.
- ♦ Develop the infrastructure necessary to have valid and reliable performance information and use it to make decisions. This would build on the current budgeting, self-assessment, and peer review systems and be based on an understanding of best practices in performance assessment. This infrastructure includes increased knowledge, skill, and networks in performance measurement and evaluation. It also includes integration of the budgeting, planning, program execution, and evaluation decision processes.

## References

- Alles, D.S. 1991. U.S. Patent Productivity. *Research Management* XXIX(5):29-35.
- American Society for Quality. 2000. American National Standard: Quality Guidelines for Research. ANSI/ASQ Z1.13-1999. Available URL: <http://www.asq.org>.
- Anderson, Bjorn and Per-Gaute Pettersen. 1996. *The Benchmarking Handbook: Step-by-Step Instructions*. London: Chapman & Hall.
- Averch, H.A. 1994. Economic Approaches to the Evaluation of Research. *Evaluation Review*, 18:77-88.
- Bendell, Tony, Louise Boulter, and Kerry Gatford. 1997. *The Benchmarking Workout: A Toolkit to Help You Construct a World Class Organization*. London: FT Pitman Publishing.

- Bozeman, B. 1993. Peer Review and Evaluation of R&D Impacts. In *Evaluating R&D Impacts: Methods and Practice*. B. Bozeman and J. Melkers (eds.). Kluwer Academic Publishers: Norwell, MA.
- Bozeman, B., and H.K. Klein. 1999. The Case Study as a Research Heuristic: Lessons from the R&D Value Mapping Project. *Evaluation and Program Planning* 22:91-103.
- Bozeman, B. and Melkers, J., (eds.). 1993. *Evaluating R&D Impacts: Methods and Practice*. Norwell, MA: Kluwer Academic Publishers.
- Bozeman, B., J.S. Dietz, and M. Gaughan. 1999. *Scientific and Technical Human Capital: An Alternative Model for Research Evaluation*. Available URL <http://rvm.pp.gatech.edu/papers>.
- Branch, Kristi M., Melissa Peffers, Rosalie Ruegg, and Robert Vallario. 2001. *The Science Manager's Resource Guide to Case Studies*. Richland, Washington: Pacific Northwest National Laboratory.
- Brown, E. 1996. Conforming the Government R&D Function with the Requirements of the Government Performance and Results Act. *Scientometrics* 36:445-470
- Brown, M.G. 1996. Keeping Score: Using the Right Metrics to Drive World-Class Performance. New York: Quality Resources.
- Callon, M., P. Larédo and P. Mustar. 1997. *The Strategic Management of Research and Technology*. Paris: Economica International.
- Cameron, Kim S., and Robert E. Quinn. 1999. *Diagnosing and Changing Organizational Culture: Based on the Competing Values Framework*. Reading, MA: Addison-Wesley Publishing Co.
- Camp, Robert C. 1995. *Business Process Benchmarking: Finding and Implementing Best Practices*. Milwaukee, Wisconsin: ASQC Quality Press.
- Camp, Robert C. (ed.). 1998. *Global Cases in Benchmarking: Best Practices from Organizations Around the World*. Milwaukee, Wisconsin: ASQ Quality Press.
- Canada, Auditor General. 1994. Science and Technology: Overall Management of Federal Science and Technology Activities. *1994 Report of the Auditor General of Canada*, Chapters 9 and 10. Ottawa.
- Canada, Federal Government. 1993. Methods for Assessing the Socioeconomic Impacts of Government S&T. Working Group on S&T Financial Management and Mechanisms. Ottawa.
- Chubin, D.E. 1994. Grants Peer Review in Theory and Practice. *Evaluation Review* 18:20-30.
- COSEPUP (Committee on Science, Engineering, and Public Policy), National Academy of Sciences. 2000. *Experiments in International Benchmarking of U.S. Research Fields*. Washington, DC: National Academy Press
- COSEPUP (Committee on Science, Engineering, and Public Policy), National Academy of Sciences. 1999. *Evaluating Federal Research Programs: Research and the Government Performance and Results Act*. Washington, DC: National Academy Press.
- Cozzens, S.E. 1999. Are New Accountability Rules Bad for Science? *Issues in Science and Technology* Summer:59-66.
- Cozzens, S.E. 1997. The Knowledge Pool: Measurement Challenges in Evaluating Fundamental Research Programs, *Evaluation and Program Planning* 20(1):77-89.

- Cozzens, S., S. Popper, J. Bonomo, K. Koizumi, and A. Flannagan. 1994. *Methods for Evaluating Fundamental Science*. DRU-875/2-CTI. RAND.
- Ellis, L. 1997. *Evaluation of R&D Processes: Effectiveness through Measurements*. Boston: Artech House.
- Feller, I. 2000. *The Who, What, ... and How of Evaluating Science and Technology Programs*, Presented at the U.S./European Workshop on S&T Policy Evaluation. Sponsored by the School of Public Policy at Georgia Institute of Technology and the Fraunhofer Institute for Systems and Innovation Research (ISI), Germany, September. <<http://www.cherry.gatech.edu/e-value>>
- Fuller, S. 1997. *Science*. Minneapolis: University of Minnesota Press.
- GAO (General Accounting Office). 2000. Government Performance and Results Act: Information on the National Science Foundation's Performance Report for the Fiscal Year 1999 and Performance Plans for Fiscal Years 2000 and 2001. GAO/RCED-00-281R.
- GAO (General Accounting Office). 1999. Performance Plans: Selected Approaches for Verification and Validation of Agency Performance Information. GGD-99-139.
- GAO (General Accounting Office). 1997. *Measuring Performance: Strengths and Limitations of Research Indicators*. GAO/RCED-97-91. Washington, DC: GAO.
- GAO (General Accounting Office). 1996. Executive Guide: Effectively Implementing the Government Performance and Results Act. GAO/GGD-96-118.
- Geisler, E. 2000. *The Metrics of Science and Technology*. Westport, CT: Quorum Books.
- Geisler, E. 1999. *The Metrics of Technology Evaluation: Where We Stand and Where We Should Go from Here*. Presented at the 24<sup>th</sup> Annual Technology Transfer Society Meeting.
- Georghiou, L. and D. Roessner. 2000. Evaluating Technology Programs: Tools and Methods. *Research Policy* 29:657-678.
- Guy, K. 2000. *Assessing Programme Portfolios vs Multi-Module Approaches*. Presented at the U.S./European Workshop on S&T Policy Evaluation. Sponsored by the School of Public Policy at Georgia Institute of Technology and the Fraunhofer Institute for Systems and Innovation Research (ISI). Germany. September. <<http://www.cherry.gatech.edu/e-value>>
- Hatry, H.P. 1999. *Performance Measurement: Getting Results*. Washington, DC: Urban Institute Press.
- House, P.W., and R.D. Schull. 1985. *Managing Research on Demand*. Lanham, MD: Abt Associates and University Press of America.
- IIT (Illinois Institute of Technology) Research Institute. 1968. *Technology in Retrospect and Critical Events in Science*. Washington, DC: National Science Foundation.
- Jordan, G.B., and L.D. Streit. 2000. *Recognizing the Competing Values in Science and Technology Organizations: Implications for Evaluation*. Presented at the U.S./European Workshop on S&T Policy Evaluation. Sponsored by the School of Public Policy at Georgia Institute of Technology and the Fraunhofer Institute for Systems and Innovation Research (ISI). Germany. September. <<http://www.cherry.gatech.edu/e-value>>
- Kaplan, R. S., and D. P. Norton. 1996. Using the Balanced Scorecard as a Strategic Management System. *Harvard Business Review* 74:75-85. <http://www.hbsp.harvard.edu/frames/groups/hbr/janfeb96/96107.html>

- Kingsley, G. 1993. The Use of Case Studies in R&D Impact Evaluations. In *Evaluating R&D Impacts: Methods and Practice*. B. Bozeman and J. Melkers (eds). Norwell, MA: Kluwer..
- Kostoff, R. N. 1997. The Principles and Practices of Peer Review. *Science and Engineering Ethics* (Special Issue on Peer Review) 3:19-34.
- Laredo, P., and P. Mustar. 2000. Laboratory Activity Profiles: An Exploratory Approach. *Scientometrics* (in publication).
- Mansfield, E. 1991. Social Returns from R&D: Findings, Methods and Limitations. *Research-Technology Management* (November-December):24-27.
- NPR (National Partnership for Reinventing Government). 1999. *Balancing Measures: Best Practices in Performance Management*. Washington, DC: Government Printing Office.
- NPR (National Partnership for Reinventing Government). 1997. *Serving the American Public: Best Practices in Performance Measurement. Benchmarking Study Report*. Washington, DC: Government Printing Office.
- NSF (National Science Foundation). 2001. *Research Assessment: What's Next (2001)*, Workshop Proceedings, Sponsored by the NSF Program of Research on Science and Technology, Airlie House Conference Center, outside Washington, DC. May, <<http://www.reseval.net>>
- NSF (National Science Foundation). 2000. *NSF GPRA Strategic Plan for FY 2001-2006*. Available URL: <http://www.nsf.gov/pubs/2001/nsf0104/nsf0104.htm>.
- NSTC (National Science and Technology Council), Committee on Fundamental Science. 1996. *Assessing Fundamental Science*. Available URL: <http://www.nsf.gov/sbe/srs/ostp/assess/start.htm>
- OECD (Organization for Economic Co-operation and Development). 1997. *In Search of Results: Performance Management Practices*.
- OMB (Office of Management and Budget). 1999. Preparation and Submission of Strategic Plans, Annual Performance Plans, and Annual Program Performance Reports. Circular A-11, part two.
- Pacific Northwest National Laboratory. 1992. Measuring Investment in R&D Excellence: A Study of Laboratory Directed Research and Development at the Nine DOE Multiprogram Laboratories. Draft Report. Washington, DC: PNNL.
- Popper, S. 1995. Economic Approaches to Measuring the Performance and Benefits of Fundamental Science. RAND PM-409-OSTP. July.
- Ransley, D.L. 1994. Do's and Don'ts of R&D Benchmarking. *Research-Technology Management* 37:50-56.
- Raytheon. 1998. *Benchmarking and Best Practice Sharing*. A Raytheon Pamphlet #M86063; RA-20020.
- Raytheon. n.d. Metrics: A Management Guide for the Development and Deployment of Strategic Metrics. A Raytheon Pamphlet #RA29907.
- Rip, A. 2000. *Societal Challenges to Evaluation*. Presented at the U.S./European Workshop on S&T Policy Evaluation. Sponsored by the School of Public Policy at Georgia Institute of Technology and the Fraunhofer Institute for Systems and Innovation Research (ISI). Germany. September. <<http://www.cherry.gatech.edu/e-value>>

- Rogers, J.D., and B. Bozeman. 2000. *Knowledge Value Alliances: An Alternative to R&D Project Evaluation*. Available URL: <http://rvm.pp.gatech.edu/papers/newtaxon.doc>
- Schumann, Jr., P.A., D.L. Ransley, and D.C.L. Prestwood. 1995. Measuring R&D Performance. *Research-Technology Management* 38:45-54.
- Sherwin, C.W., and R.S. Isenson. 1967. Project Hindsight: Defense Department Study of the Utility of Research. *Science* 156:1571-1577.
- Szakonyi, R. 1994. Measuring R&D Effectiveness I. *Research-Technology Management* 37:27-33.
- Tassey, G. 1996. Rates of Return from Investments in Technology Infrastructure. NIST Planning Report 96-3.
- Tipping, J.W., E. Zeffren, and A.R. Fusfeld. 1995. Assessing the Value of Your Technology. *Research- Technology Management* 38:22-39. (See update at [iriinc.org](http://iriinc.org).)
- U.S. Congress, Office of Technology Assessment (OTA). 1986. *Research Funding as an Investment: Can We Measure the Returns?* OTA-TM-SET-36. Washington, DC: U.S. Government Printing Office.
- U. S. Congress. 1993. Government Performance and Results Act of 1993. PL 103-62.
- U.S. Federal Government. 1994. *Science in the National Interest*. Fundamental Science Report. Available URL: [http://www.whitehouse.gov/WH/EOP/OSTP/Science/html/Sitni\\_Home.html](http://www.whitehouse.gov/WH/EOP/OSTP/Science/html/Sitni_Home.html).
- University of California. 1995. *University Laboratory Self-Assessment and Annual Review Manual*. Appendix F, Rev. 2, June. Available URL: <http://www.llnl.gov>.
- Vonortas, N., and M. Lackey. 2000. *Real Options for Public Sector Investments*. Presented at the U.S./European Workshop on S&T Policy Evaluation. Sponsored by the School of Public Policy at Georgia Institute of Technology and the Fraunhofer Institute for Systems and Innovation Research (ISI), Germany, September. <<http://www.cherry.gatech.edu/e-value>>
- Yin, Robert K. 1984. *Case Study Research: Design and Methods*. 2<sup>nd</sup> Edition. Thousand Oaks, CA: Sage.
- Werner, B.M., and W.E. Souder. 1997. Measuring R&D Performance – State of the Art. *Research-Technology Management* 40:34-42.